
UCRL-TR-202878

# Stochastic Engine Final Report:

# Applying Markov Chain Monte Carlo Methods with Importance Sampling to Large-Scale Data-Driven Simulation

*R. D. Aines, J. J. Nitao, W. G. Hanley, S. Carle, A. L. Ramirez, R. L. Newmark, V. M. Johnson, R. E. Glaser, S. Sengupta, B. Kosovic, K. M. Dyer, K. A. Henderson, G. A. Sugiyama, T. L. Hickling, G.A. Franz, M. E. Pasyanos, D. A. Jones, R. J. Grimm, G. Johannesson, and R. A. Levine*

**March 2004**

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

# Stochastic Engine Final Report: Applying Markov Chain Monte Carlo Methods with Importance Sampling to Large-Scale Data-Driven Simulation

Roger D. Aines, John J. Nitao, William G. Hanley, Steven Carle, Abelardo L. Ramirez, Robin L. Newmark, Virginia M. Johnson, Ronald E. Glaser, Sailes Sengupta, Branko Kosovic, Kathleen M. Dyer, Keith A. Henderson, Gayle A. Sugiyama, Tracy L. Hickling, G.A. Franz, Michael E. Pasyanos, Dale A. Jones, Roger J. Grimm, and Gardar Johannesson,
*Lawrence Livermore Nat'l Lab*

Richard A. Levine
*San Diego State University*

## *Abstract*

Accurate prediction of complex phenomena can be greatly enhanced through the use of data and observations to update simulations. The ability to create these data-driven simulations is limited by error and uncertainty in both the data and the simulation. The stochastic engine project addressed this problem through the development and application of a family of Markov Chain Monte Carlo methods utilizing importance sampling driven by forward simulators to minimize time spent search very large state spaces. The stochastic engine rapidly chooses among a very large number of hypothesized states and selects those that are consistent (within error) with all the information at hand. Predicted measurements from the simulator are used to estimate the likelihood of actual measurements, which in turn reduces the uncertainty in the original sample space via a conditional probability method called Bayesian inferencing. This highly efficient, staged Metropolis-type search algorithm allows us to address extremely complex problems and opens the door to solving many data-driven, nonlinear, multidimensional problems.

A key challenge has been developing representation methods that integrate the local details of real data with the global physics of the simulations, enabling supercomputers to efficiently solve the problem. Development focused on large-scale problems, and on examining the mathematical robustness of the approach in diverse applications. Multiple data types were combined with large-scale simulations to evaluate systems with ~10^20,000 possible states (detecting underground leaks at the Hanford waste tanks). The probable uses of chemical process facilities were assessed using an evidence-tree representation and in-process updating. Other applications included contaminant flow paths at the Savannah River Site, locating structural flaws in buildings, improving models for seismic travel times systems used to monitor nuclear proliferation, characterizing the source of indistinct atmospheric plumes, and improving flash radiography. In the course of developing these applications, we also developed new methods to cluster and analyze the results of the state-space searches, as well as a number of algorithms to improve the search speed and efficiency. Our generalized solution contributes both a means to make more informed predictions of the behavior of very complex systems, and to improve those predictions as events unfold, using new data in real time.

## *Introduction*

How can we understand a system that is too complex to sample or impossible to observe directly, but for which we have good models that will predict behavior under specific conditions? Such systems are a large part of the LLNL mission. The stochastic engine approach addresses these problems by integrating the general knowledge represented by models, with specific knowledge represented by data. Stochastic methods have been available for many years, but have been limited in their application by the required computational resources. We have demonstrated that with modern computational systems, it is possible to apply stochastic approaches to systems that incorporate complex simulations over very large domains. This report summarizes the stochastic engine approach and lists the publications providing details of its development and application.

Determining the properties of an object that we cannot directly observe is a fundamental problem in many fields of study. Examples include study of the deep earth, intelligence activities, or attempting to reconstruct an event that happened in the past. There is always a lack of data, and yet there are often many types of data available. We use inference and models to extrapolate or interpolate our knowledge, but we are fundamentally limited by the inability to inquire over the entire spatial or temporal domain of interest. There is never enough data for the most complex situations. The purpose of the stochastic engine initiative was to attack this problem by developing a method to simultaneously use many kinds of data to refine our understanding of complex systems, focusing on geologic systems. We developed and demonstrated a family of Markov chain Monte Carlo methods utilizing importance sampling driven by forward simulators to minimize time spent search very large state spaces. The stochastic engine rapidly chooses among a very large number of hypothesized states and selects those that are consistent (within error) with all the information at hand. Predicted measurements from the simulator are used to estimate the likelihood of actual measurements, which in turn reduces the uncertainty in the original sample space via conditional probability (Bayesian inferencing). This highly efficient, staged Metropolis-type search algorithm allows us to address extremely complex problems and opens the door to solving many data-driven, nonlinear, multidimensional problems.

The stochastic engine approach focuses on improving one “base” set of data (or representation of the system), from which other parameters of interest can be calculated using process models. The lithology (the general physical characteristics of a rock) of an underground system is the base representation for geological systems. It provides a ready means to predict the behavior of the system under forcing events such as injection of a fluid; when we know the lithology more accurately, we can predict the behavior of the system more accurately. The response of the system to these forcing events can then be measured to further improve the knowledge of the system. This feedback is central to our

ability to acquire enough knowledge about complicated systems; we need to utilize each layer of knowledge to improve our acquisition of new data, continuously improving the detail and accuracy of our system knowledge. The stochastic engine is designed to incorporate everything from the geologists' first field observations to the millions of measurements made during a field operation such as a steam remediation project, into an integrated and continuously improving understanding of the base representation. While our first application is geologic, this method is broadly applicable to any topical area in which direct observations of a system can be combined with general understanding represented by simulation.

Analysis of this kind is needed by all large-scale subsurface efforts. The most obvious, such as oil recovery, are those in which the cost of additional wells is very large and the goal is to maximize recovery per dollar invested. In government applications such as nuclear waste disposal or environmental remediation, a more immediate goal is often the reduction of uncertainty in the outcome of costly or long-term efforts. These data-rich applications are in contrast to data-poor situations, such as locating underground structures. In all cases the stochastic engine can improve the value of existing data, and guide the acquisition of future data through quantitative evaluation of method, location, and number of points.

The stochastic engine approach honors all data and model information to produce probability distributions identifying likely system configurations or behavior, and quantifying the potential improvement provided by new data. Even when conventional inversion and analysis methods are able to address complex problems, they provide only a single "best" answer, throwing away much of the information and precluding other likely possibilities. This hinders the subsequent use of the analysis by failing to allow for alternative possible outcomes.

## *Approach*

The stochastic engine uses existing simulators to predict data values that are then compared to exactly analogous measurements to determine which possible configurations of a system are in fact closest to the real condition. An extremely efficient search algorithm, derived from the Metropolis/Hastings method, is used to determine which states to test. Different types of data (and their accompanying simulation) can be combined in stages, so that extremely complex state spaces can be searched quickly. The initial state space is described by a mathematical model called the base representation that includes all the salient features of the system, while being as simple as possible. This constitutes the "prior" distribution in the Bayesian inferencing scheme. The results of the engine analysis are in terms of these same states; the "posterior" distribution resulting from the analysis is the set of states that are consistent with the data and inherent error in the system.

The simulators used are all forward models, that is, they predict a value given an initial condition; these can be used with extremely non-linear problems which are difficult or impossible to directly invert. Modern computational power makes this reasonable for complicated problems, and the search algorithm has proven to be so efficient that many problems of geologic interest are tractable on workstations alone. The wide range of applicable problems is a function of the number of good models (simulators) that exist today. The stochastic engine methodology can use any model that predicts results based on initial conditions. Initial development focuses on earth-sciences models for lithology, flow and transport, geochemistry, and geophysical imaging.

Contributing data can be of many types, ranging from distinct physical or chemical measurements for which sensitivity, resolution, etc., are known, to "soft" data such as expert inference or qualitative models. The experimental methods used include multiple simultaneous imaging methods and "active" analyses such as pump tests or deformation tests that force changes in the system that are predictable if the internal structure is known. The resulting analysis is unique both in the simultaneous use of multiple data types (for instance x-ray tomography and positron emission tomography) and in the calculation of the structure directly in terms of probabilities. Rather than a single "best" structure, the stochastic engine generates a range of plausible structures and the corresponding probabilities that they are correct. This facilitates decision analysis and needs-based experimental planning.

Our development activities focused on geological/ geophysical systems for which extensive observations can be made on time scales shorter than the characteristic scale for the problem, enabling predictive understanding to evolve much faster than the real-time evolution of the natural system. These include remediation of groundwater, atmospheric transport, and characterization of natural environments. We have good process models and statistical means of describing initial conditions in these systems. While our experience with earth systems drove the initial stochastic engine development, we are using the experience from these systems to expand application to other support areas including intelligence and defense. These include WMD facility sampling strategies and evaluation, imaging and non-destructive evaluation of complex assemblies, measuring structural response to deformation forces, and locating the source of atmospheric plumes.

## *Engine Function: Integration of MCMC with Forward Simulators*

The actual connection of a hypothesis to an observation is made via a forward model: for a possible subsurface configuration the forward model predicts the values that would be

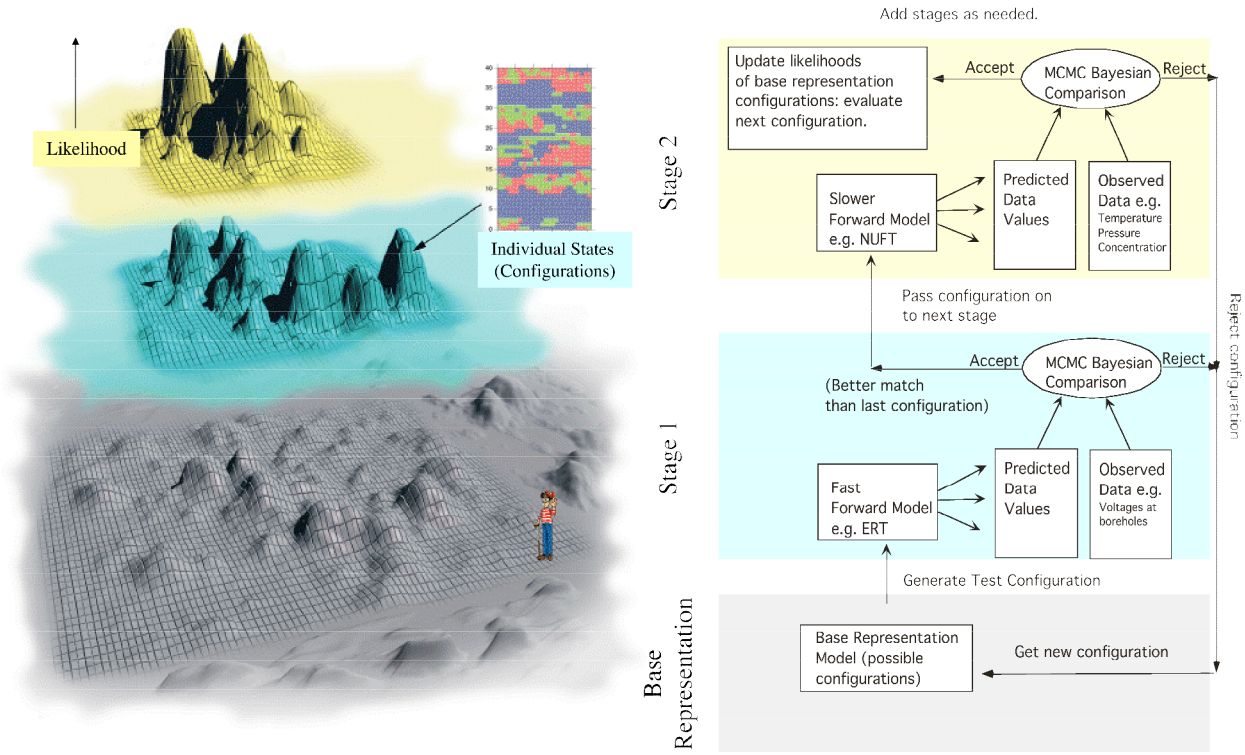


**Figure 1. The MCMC Stochastic Engine Combines Observations and Simulations To Determine The Likelihoods of Possible System Configurations**

observed by actual measurement. These are compared to real data. The degree of match between the real and predicted data is fed back to a Markov Chain Monte Carlo (MCMC) algorithm that samples candidate lithologic configurations for testing. Accepted states constitute samples from the posterior distribution and provide the basis for subsequent inference. By staging these comparisons in a series of MCMC algorithms, we can identify probable configurations using fast models early in the process. The most computationally intensive models are only used at later stages on configurations that are already known to be consistent with data used at earlier stages (Figure 1).

Configurations that pass all the stages are possible true configurations of the system. It is usually the case that the MCMC approach produces a reduction in the number of possible configurations (represented in the prior distribution) by many orders of magnitude. The approach also provides a seamless methodology for combining observational data with our forward models to produce state estimates and corresponding uncertainties that are

not readily available through conventional inversion approaches. This is an extremely powerful method for incorporating previously known information and newly acquired observations into an estimate of the probability distribution of the states of the system. By generating likelihoods of actual lithologies, we can readily involve a variety of data types in the inference process and use the obtained lithologic posterior distribution to guide further investigations. By combining configurations into meta-classes (configurations that are so similar as to behave identically in the field, within error) we can readily deduce whether there is more than one highly probable configuration for our system, and which data will be the most useful in resolving between competing configurations.

The MCMC staged algorithm is well suited to a number of improvements that we anticipate will be crucial to dealing with complex, three-dimensional problems.

- Any number of stages can be used, involving all the data available for the system.
- Initial constraints placed on the base model confine the analysis to solutions that are known to be physically realistic, speeding the search and enhancing the usefulness of the answer.
- Data can be added to the algorithm sequentially, as it becomes available.
- The algorithm can be stopped when all available information has been processed, and the newly obtained distribution of the possible configurations can then be the basis for processing new data in a subsequent staged MCMC algorithm.
- Resolution in the base model can vary across the state space, allowing focus on critical areas. Individual spatial volumes can be analyzed by their own stages, and the result collapsed to a single probability distribution (as described above).
- Additional parameters can be added to the analysis by mapping them onto the lithologic representation. For instance, the presence of contaminant can be added as a representation element and a series of stages for chemical data types can be incorporated into the simulation to resolve the location of the contaminant.

## *Results*

A number of reports have been prepared describing the development of stochastic methods for the engine, and the use of the method in analyzing real problems. The full method was described previously:

UCRL-ID-148221 (2002) *The Stochastic Engine Initiative: Improving Prediction of Behavior in Geologic Environments We Cannot Directly Observe* Aines, R.D., J.J. Nitao, R.L. Newmark, S. Carle, A.L. Ramirez, D.B. Harris, J.W. Johnson, V.M. Johsnon, D. Ermak, G. Sugiyama, W.G. Hanley, S. Sengupta, W. Daily, R. Glaser, K. Dyer, G Fogg, Y. Zhang, Z. Yu, R. Levine

A series of more detailed papers describing individual components and applications was prepared at the conclusion of the project. The abstracts and references are given below.

Base Representations and Analysis

The base representation of lithology used in the most complex stochastic engine applications is based on the code TSIM.

> Steven Carle
> Integration of Soft Data in Categorical Geostatistical Simulation UCRL-ID-153653 (rev2)
>
> Abstract
>
> Abundant uncertain or indirect "soft data" on lithology (e.g., geophysical logs, drillers' logs, etc.) offer potential constraints for subsurface models. Previous geostatistical simulation algorithms have not fully addressed the impactof data uncertainty in formulation of (co)kriging equations and simulated annealing objective functions. This paper introduces the categorical geostatistical simulation code tsim-s, which accounts for indicator data uncertainty through a data "hardness" parameter. In generating geostatistical realizations with tsim-s, the impact of uncertainty in the soft data is factored into formulation of both the cokriging and simulated quenching steps of the simulation algorithm. The degree to which soft data reduces variability in simulation outputs is quantified by mapping category probabilities derived by averaging indicator values from many realizations. In addition to point or borehole data, arrays of data, including other realizations, can also be used as soft conditioning. Future applications of soft conditioning in geostatistical simulation may enable manipulation of heterogeneity structure in flow and transport model calibration, stochastic inverse approaches, or sensitivity analysis.

The result of a stochastic engine analysis is a set of states that have been accepted by the analysis. The sum of these states, which may be regarded as a type of stacking of the states, is the full posterior probability distribution. Unfortunately this may still encompass too much information to be immediately accessible to the analyst. One approach is to combine all the similar states into groups called clusters.

> Sailes K. Sengupta and Abelardo L. Ramirez
> Exploratory global inference from posterior image samples in a Markov Chain Monte Carlo simulation experiment: A high-dimensional data clustering approach with extensions to categorical data UCRL-JRNL-200985
>
> Abstract
>
> Global inference directly from images obtained as posterior samples in a simulation experiment such as Markov Chain Monte Carlo may be possible by a successful clustering in the space of images as *meta-states*. This paper explores the classical algorithms such as K-Means and ISODATA and extensions thereof to achieve this. It also explores the possibilities of modifying the K-means clustering algorithm due to Hartigan in order to accommodate categorical data. We use the nearest neighbor clustering principle like in the K-means algorithm. However, when clustering categorical data, two problems must be addressed. First, the distance between two samples has to be suitably defined. Second, since the cluster centroids are no longer definable, we must have an appropriate measure that gives the distance between a sample and a cluster. The first is

achieved by defining an extended form of Hamming distance. The second is accomplished by defining a new sample-cluster distance based on the pixel- (component) wise histogram array for each cluster

## Stochastic Search Algorithms

The efficiency of the stochastic approach is dependent upon the speed with which the algorithm traverses the state space. Several improvements to the standard approaches were made:

Levine, RA, Yu, Z, Hanley, W.G., Nitao, J.J.
Adaptive Scanning Strategies: The Hastings Sampler UCRL-JC-152437

Abstract
Markov chain Monte Carlo (MCMC) routines have revolutionized the application of Monte Carlo methods in statistical application and statistical computing methodology. The Hastings sampler, encompassing both the Gibbs and Metropolis samplers as special cases, is the most commonly applied MCMC algorithm. The performance of the Hastings sampler relies heavily on the choice of sweep strategy, that is, the method by which the components or blocks of the random variable X of interest are visited and updated, and the choice of proposal distribution, that is the distribution from which candidate variates be drawn for the accept-reject rule each iteration of the algorithm. We focus on the random sweep strategy, where the components of X are updated in a random order, and random proposal distributions, where the proposal distribution is characterized by a randomly generated parameter. We develop an adaptive Hastings sampler which learns from and adapts to random variates generated during the algorithm towards choosing the optimal random sweep strategy and proposal distribution for the prob- lem at hand. As part of the development, we prove convergence of the random variates to the distribution of interest and discuss practical implementations of the methods. We illustrate the results presented by applying the adaptive componentwise Hastings samplers developed to sample multivariate Gaussian target distributions and Markov random field models.

Levine, RA, Yu, Z, Hanley, W.G., Nitao, J.J.
Adaptive Scanning Strategies: The Gibbs Sampler UCRL-JC-152443

Abstract The Gibbs sampler, being a popular routine amongst Markov chain Monte Carlo sampling methodologies, has revolutionized the application of Monte Carlo methods in statistical computing practice. The performance of the Gibbs sampler relies heavily on the choice of sweep strategy, that is, the means by which the components or blocks of the random vector X of interest are visite d and updated. We develop an automated, adaptive algorithm for implementing the optimal sweep strategy as the Gibbs sampler traverses the sample space. The decision rules through which this strategy is chosen are based on convergence properties of the induced chain and precision of statistical inferences drawn from the generated Monte Carlo samples. As part of the development, we analytically derive closed form expressions for the decision criteria of interest and present computationaily feasible implementations of the adaptive random scan Gibbs sampler via a Gaussian approximation to the target distribution, We illustrate the results and

algorithms presented by using the adaptive random scan Gibbs sampler developed to sample multivariate Gaussian target distributions and analyze genetic linkage, s~reening test, and image data.

## Applications

Three major applications were demonstrated and published: underground imaging, determining flaws in systems such as buildings by analyzing the vibrational signature, and refining estimates of the probable uses of suspect industrial facilities by integrating disparate data types.

### Imaging:

Ramirez, A. L., J.J. Nitao, W.G. Hanley, R.D. Aines, R.E. Glaser, S.K. Sengupta, K.M. Dyer, T.L. Hickling, W.D. Daily
Stochastic Inversion of Electrical Resistivity Changes Using a Markov Chain, Monte Carlo Approach UCRL-JC-155048 Rev. 1

Abstract

We describe a stochastic inversion method for mapping subsurface regions where the electrical resistivity is changing. The method combines prior information, electrical resistance data and forward finite difference models to produce subsurface resistivity models that are most consistent with all available data. Bayesian inference and a Metropolis simulation algorithm form the basis for this approach. The approach enables the estimation of distributions of both individual parameters such as center of mass and groups of parameters such as resistivity change images. Attractive features of this approach include its ability to: 1) provide quantitative measures of the uncertainty of a generated estimate, 2) seamlessly integrate disparate data such as electrical resistance measurements and liquid volume measurements, 3) effectively invert complex nonlinear forward models without appealing to unrealistic simplifying assumptions, 4) function effectively when exposed to degraded conditions including: noisy data, incomplete data sets and model misspecification and, 5) allow alternative model estimates to be identified, compared and ranked. The proposed method is computationally expensive, requiring the use of large computer clusters to make its application practical. A series of physical model test cases have been performed to validate the methodology. Field results using data collected during the infiltration of a salt-water tracer are also discussed. Methods that assess MCMC convergence and summarize interesting features of the posterior distribution are introduced. The stochastic inversions presented suggest that zones of resistivity change can be successfully mapped with this approach. The stochastic tomographs accurately identify the most probable location, shape, and volume of the changing region, and the most likely resistivity change.

### Suspect Facility Evaluation:

Virginia M. Johnson Tracy L. Hickling John J. Nitao Dale A. Jones Roger J. Grimm
A Stochastic Approach to Assessing Chemical Production Activities at an Industrial Site: A National Security Application of the Stochastic Engine Initiative UCRL-TR-200691 (This document is OUO)

Summary for unrestricted use:

A problem that plagues many national security data analysis applications is an absence of tools for weighing alternative hypotheses regarding a problem. Suppose, for example, that an industrial facility is being monitored for potential involvement in a chemical weapons program. It may be producing agent A, producing some legitimate compound B, or engaging in both activities. Although we can enumerate the alternatives, it is currently difficult, at least in an objective way, to make statements such as "There is a 75% probability that agent A is being manufactured at the site" or "There is a 50% chance that both agent A and compound B are being produced". The ability to assign probabilities to alternative hypotheses would give considerably more guidance to decision-makers who must risk the consequences of being wrong. This report describes the current status of a project to apply stochastic methods to the problem of estimating the likelihood of various hypotheses concerning potential production activity at industrial sites.

## Structure Flaw Analysis

A Markov Chain Monte Carlo Based Method for System Identification
R. E. Glaser, W. G. Hanley, C. L. Lee, J. J. Nitao UCRL-ID-150494

Abstract

This paper describes a methodology for the identification of mechanical systems and structures from vibration response measurements. It combines prior information, observational data and predictive finite element models to produce configurations and system parameter values that are most consistent with the available data and model. Bayesian inference and a Metropolis-Hastings simulation algorithm form the basis for this approach. The resulting process enables the estimation of distributions of both individual parameters and system-wide states. Attractive features of this approach include its ability to: 1) provide quantitative measures of the uncertainty of a generated estimate; 2) function effectively when exposed to degraded conditions including: noisy data, incomplete data sets and model misspecification; 3) allow alternative estimates to be produced and compared; and 4) incrementally update initial estimates and analysis as more data become available. A series of test cases based on a simple fixed-free cantilever beam is presented. These results demonstrate that the algorithm is able to identify the system, based on the stiffness matrix, given applied force and resultant nodal displacements. Moreover, it effectively identifies locations on the beam where damage (represented by a change in elastic modulus) was specified.

## Atmospheric Inversion

Sensor-Driven Modeling Capability for Determining Unknown Source Terms
Branko Kosovic and Michael Sohn, UCRL-JC-153404

This extended abstract describes the goals and initial of the atmospheric inversion application. This work is continuing under separate LDRD development led by Gayle Sugiyama.

Software

The software written to perform these applications is a family of Python, C, C++, and FORTRAN codes linked by a Python backbone. For the large simulation codes such as NUFT, a Python wrapper was written to incorporate the code into the stochastic engine package without any changes to the original software. The overall architecture, coding standards, and methods for incorporating new codes (applications) into the stochastic engine package are described in the software development manual:

> Franz, A., K. Dyer, K. Henderson, R. Aines, J. Nitao, A. Ramirez, R. Glaser, W. Hanley Stochastic Engine Software Development Guide UCRL-TM-200386

## *Acknowledgement*

This research was funded by the Laboratory Directed Research and Development (LDRD) Program at Lawrence Livermore National Laboratory (LLNL). The LDRD Program is mandated by Congress to fund director-initiated, long-term research and development (R&D) projects in support of the DOE and national laboratories mission areas. The Director's Office LDRD Program at LLNL funds creative and innovative R&D to ensure the scientific vitality of the Laboratory in mission-related scientific disciplines.